УДК 519.767

МОДЕЛЬ РАЗБИЕНИЯ МНОЖЕСТВА ЭЛЕМЕНТОВ СМЫСЛА МНОГОЗНАЧНЫХ СЛОВ ПЕРЕВОДИМОГО ПРЕДЛОЖЕНИЯ В СИСТЕМАХ АВТОМАТИЧЕСКОГО ПЕРЕВОДА

Н.Ф. Хайрова[1], Н.В. Шаронова[2]

[1] ХГУ «НУА», г. Харьков, Украина, nikhayv@vlink.kharkov.ua

[2] НТУ «ХПИ» г. Харьков, Украина, sharonova@kpi.kharkov.ua

Проведен анализ задач семантического анализа систем машинного перевода. Разработана модель снятия семантической омонимии многозначных слов переводимого предложения. Для разбиения множества элементов смысла многозначных слов используется метод компараторной идентификации. Показаны преимущества использования данной модели для снятия семантической многозначности в процессе автоматического перевода.

МАШИННЫЙ ПЕРЕВОД, СЕМАНТИЧЕСКИЙ АНАЛИЗ, КОМПАРАТОРНАЯ ИДЕНТИФИКАЦИЯ, СЕМАНТИЧЕСКАЯ ОМОНИМИЯ, КОМПОНЕНТНЫЙ АНАЛИЗ

Введение

Среди большого числа направлений интеллектуальной деятельности человека, срочно требующих автоматизации, значительное место в современном мире занимает наиболее сложный вид этой деятельности — перевод с одного естественного языка на другой естественный язык.

Потребность в автоматизации процесса перевода возникла уже давно, достаточно сказать, что первая программа машинного перевода была продемонстрирована в 1954 году, и с тех пор более 50 лет в различных научных школах мира проводятся научные исследования по улучшению качества такого перевода, главными преимуществами которого остаются скорость и дешевизна.

Одной из главных составляющих системы машинного перевода, от которой непосредственно зависит качество перевода, является лингвистический процессор. Идеальный лингвистический процессор, позволяющий создать идеальную систему, представляет собой набор процедур, осуществляющих переход от исходного текста на переводимом языке к “голому” смыслу (анализ) и назад к готовому тексту на языке перевода (синтез) [1] (рис. 1).

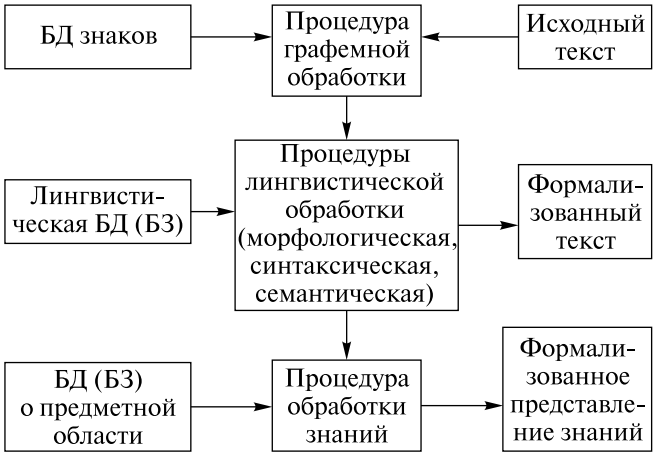

Рис.1. Схема идеального лингвистического процессора

1. Актуальность исследования

На поверхностных уровнях моделирования естественного языка в системах автоматической обработки текста — морфологическом и синтаксическом, уже достигнуто достаточно много практических результатов. В аннотациях к последним версиям систем машинного перевода обычно сообщается об отсутствии грамматических ошибок.

Этап же семантического анализа по-прежнему является одним из наиболее актуальных направлений исследований в области ИИ. И если под семантическим анализом в общем случае понимается представление значения входного текста в терминах некоторого формального языка, «понятного» ЭВМ, то в современных системах машинного перевода широкого использования сущность семантического анализа сводится к выбору нужного значения переводного эквивалента [2].

Сегодня существует несколько основных направлений снятия семантической омонимии в системах машинного перевода:

- метод семантических фильтров;
- компонентный анализ;
- семантическая обработка с использованием фреймов.

Наиболее перспективным, но по-прежнему трудно реализуемым методом, является компонентный анализ, применяемый в интерлинговых системах машинного перевода.

2. Постановка задачи исследования

При компонентном анализе весь словарь системы описывают с помощью ограниченного и сравнительно небольшого числа семантических признаков (сем). При переводе предложения семантическую омонимию снимают за счет привлечения к переводу знаний о тематике предложения, которые извлекаются из сочетаемости сем многозначных слов предложения.

В направлении развития компонентного анализа предлагается для выбора нужного значения (а значит правильного переводного эквивалента) многозначного слова использовать метод компараторной идентификации, который разработан научной школой Юрия Петровича Шабанова-Кушнаренко при решении проблем ИИ [3].

Вводим множество T переводных эквивалентов многозначных слов переводимого предложения (ПЭМСП) $T = \{t_i\}$, $1 \le i \le n$. Введем также основное для наших рассмотрений понятие сем S. Под семами понимают атомы семантического признака, множеством которых можно описать всевозможные понятия, объединяемые в многозначном слове. Понятие формируется в сфере мышления и имеет внеязыковую природу. Но поскольку мысль не может существовать вне слова, под семой мы будем понимать лексическую единицу, представляющую определенное значение слова. Введем достаточно четко очерченное множество сем многозначных слов предложения $S = \{s_j\}$, $1 \le j \le m$.

Два множества T и S являются базовыми при использовании метода компараторной идентификации переводных эквивалентов многозначных слов переводимого предложения. Центральной задачей семантического анализа предложения методом компараторной идентификации является разбиение всех сем рассматриваемых многозначных слов предложения на классы эквивалентности с тождественным или почти тождественным смыслом, то есть отнесение их к определенным темам.

3. Введение компонентно-семантического предиката

Введем бинарный компонентно-семантический предикат P, заданный на декартовом произведении множеств ПЭМСП и сем, отражающих смысл этих переводных эквивалентов.

Пусть компаратор реализует бинарный предикат $P(t)$, заданный на декартовом произведении $T \times S$ множеств T и S. Предполагается, что семы из множества S описывают в сжатом виде смысл множества ПЭМСП T. Поэтому, когда классификатор воспринимает пару (t, s), образованную из переводного эквивалента и семы, он устанавливает, соответствует ли сема данному переводному эквиваленту.

Компаратор, воспринимая пару, образованную из переводного эквивалента и семы, может установить, соответствует ли сема данному переводному эквиваленту, причем любой предикат $P(t, s)$ однозначно приравнивается 0 или 1.

Если предикат P равен 1, то это значит, что сема s соответствует переводному эквиваленту многозначного слова предложения t.

Если компаратор рассмотрел все возможные пары, то результат работы можно представить в виде двудольного графа, где верхнее множество вершин — это все ПЭМСП t_i, а нижнее — семы, отражающие смысл переводных эквивалентов s_j. Дуга проводится тогда и только тогда, когда предикат $P(t_i, s_j) = 1$. Эта система ребер описывает отображение верхнего множества в нижнее и отображение нижнего множества в верхнее.

Пример подобного графа показан на рис. 2.

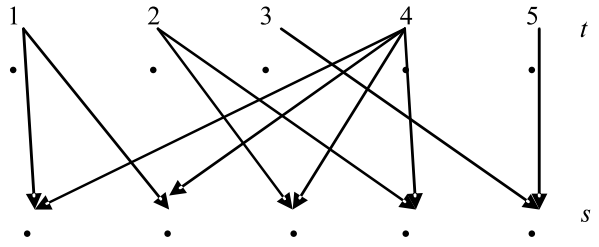

Рис. 2. Пример двудольного графа, отображающего ПЭМСП в множество сем

В этом примере множество S разбивается на три класса: $S_1 = \{s_1, s_2\}$, $S_2 = \{s_3, s_4\}$, $S_3 = \{s_5\}$.

На декартовом произведении множества ПЭМСП T можно ввести предикат эквивалентности, который отображает соответствие переводных эквивалентов одной микротеме переводимого предложения:

$$E_1(t_1, t_2) = \forall\, s \in S\,(P(t_1, s) \sim P(t_2, s)). \qquad (1)$$

Предикат E_1 является предикатом эквивалентности и однозначно определяется предикатом P. Предикат $E_1(t_1, t_2)$ можно использовать для объективного определения отношения любых двух ПЭМСП t_1 и t_2, принадлежащих множеству T, к одной микротеме. Действительно, если $E_1(t_1, t_2) = 1$, то при семе из множества S: $P(t_1, s) = P(t_2, s)$. Это означает, что элементы смысла переводных эквивалентов t_1 и t_2, выражаемые семами из множества S, совпадают, следовательно, классификатор отнесет переводные эквиваленты t_1 и t_2 к одной микротеме. Если же $E_1(t_1, t_2) = 0$, то найдется такая сема $s \in S$, для которой $P(t_1, s) \ne P(t_2, s)$. В этом случае не весь смысл переводных эквивалентов t_1 и t_2, отражаемый элементами смысла из множества S, совпадает, следовательно, два переводных эквивалента относятся к разным микротемам и не могут включаться в одно переводимое предложение.

4. Компараторная идентификация элементов смысла многозначных слов

Введенный компонентно-семантический предикат $P(t, s)$ и полученный предикат эквивалентности E_1 (1) позволяют провести разбиение ПЭМСП на классы эквивалентности, представляющие собой определенные микротемы предложения. При этом

для каждого класса можно ввести обозначение микротемы, объединяющей данный класс. Ясно, что переводные эквиваленты, входящие в полученные нами классы эквивалентностей, не тождественны по смыслу — они являются эквивалентными относительно выражаемой ими микротемы. Предикат E_1 определяет разбиение ϑ_1 множества T на слои переводных эквивалентов переводимого предложения. Все переводные эквиваленты, принадлежащие одному слою разбиения, относятся к одной микротеме. Любые же ПЭМСП, взятые из разных слоев разбиения, относятся к различным подтемам. Ясно, что для правильного перевода предложения следует использовать переводные эквиваленты, относящиеся к одному слою разбиения.

Классу L_q всех ПЭМСП $t \in T$, относящихся к одной подтеме, содержащему переводной эквивалент $q \in T$, соответствует предикат $L_q(t) = E_1(t, q)$. Учитывая зависимость (1), получаем

$$L_q(t) = \forall s \in S(P(t, s) \sim P(q, s)). \qquad (2)$$

Формула (2) выражает деление ПЭМСП на микротемы через предикат P, объективно определяемый компаратором.

5. Пример работы компаратора при переводе английского предложения

Рассмотрим работу компаратора на примере перевода английского предложения, включающего многозначные слова: *Some common Web client interfaces — also Known as Web browsers — include NetScape, Internet Explorer and others check electronic documents* [4]. Рассматриваемое предложение включает три многозначных слова *web*, *interface* и *browser*. Множество ПЭМСП T= $\{q_1, ..., q_{11}\}$: q_1 = перепонка, q_2=паутина, q_3=сеть, q_4=соединительная ткань, q_5=соединительная конструкция, q_6 = интерфейс, q_7 = стык, q_8 = граница между двумя материалами, q_9 = животное, q_{10} = человек, перелистывающий книги, q_{11}=браузер.

Также априорно из словаря выделяется 6 сем, отображающих смысл данных переводных эквивалентов многозначных слов переводимого предложения, $S = \{l_1, ..., l_6\}$: l_1=фауна, l_2=компьютер, l_3=Интернет, l_4=анатомия, l_5=техника, l_6=химия.

Компаратор последовательно перебирает все пары переводных эквивалентов и соответствующих им сем. Если сема соответствуют данному переводному эквиваленту, предикат $P(t_i, s_j)$ $1 \le i \le 11$, $1 \le j \le 6$ обращается в 1; если же соответствие не установлено, предикат принимает значение 0. Перебрав все возможные пары (t_i, s_j), получаем компонентно-семантическим предикат $P(t, s)$, применив к которому формулу (2) получим слои разбиения переводных эквивалентов многозначных слов предложения на микротемы.

Предикат *P(t, s)* задан двудольным графом, показанным на рис. 3.

Используя предикат узнавания [3]:

$$a_i(x_j) = \begin{cases} 1, & \text{если } x_j = a_i; \\ 0, & \text{если } x_j = a_i, \end{cases} \qquad (3)$$

запишем предикат классификации P следующей формулой:

$$P(t, s) = t^{q_1}s^{l_1} \vee t^{q_2}s^{l_1} \vee t^{q_3}t^{l_2} \vee t^{q_3}s^{l_3} \vee t^{q_4}s^{l_4} \vee$$
$$\vee t^{q_5}s^{l_5} \vee t^{q_6}s^{l_2} \vee t^{q_7}s^{l_5} \vee t^{q_8}s^{l_5} \vee t^{q_8}s^{l_6} \vee t^{q_9}s^{l_1} \vee \qquad (4)$$
$$\vee t^{q_{10}}s^{l_4} \vee t^{q_{11}}s^{l_2} \vee t^{q_{11}}s^{l_3}$$

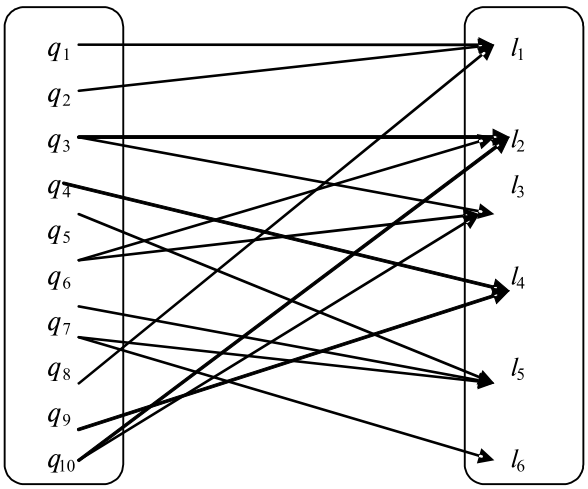

Рис. 3. Графическая интерпретация предиката $P(t, s)$

Используя формулу (2), вычисляя соответствующие предикаты, получим слои разбиения ϑ множества T:

$$L_{q1}(t) = (P(t, s_1) \sim P(q_1, s_1))\,(P(t, s_2) \sim P(q_1, s_2))\,(P(t, s_3) \sim \sim P(q_1, s_3))\,(P(t, s_4) \sim P(q_1, s_4)) \bullet (P(t, s_5) \sim P(q_1, s_5))\,(P(t, s_6) \sim \sim P(q_1, s_6))) = t^{q_1} \vee t^{q_2} \vee t^{q_9};$$
$$L_{q_2}(t) = t^{q_3} \vee t^{q_6} \vee t^{q_{11}};$$
$$L_{q_3}(t) = t^{q_3} \vee t^{q_6} \vee t^{q_{11}}; \qquad (5)$$
$$L_{q_4}(t) = t^{q_4} \vee t^{q_{10}};$$
$$L_{q_5}(t) = t^{q_5} \vee t^{q_7} \vee t^{q_8};$$
$$L_{q_6}(t) = t^{q_8}.$$

Определим повторяющиеся слои разбиения ϑ:

$$\mu = \{q_3, q_6, q_{11}\}. \qquad (6)$$

Повторяющийся слой разбиения μ представляет собой микротему, отображающую тематику переводимого предложения. Таким образом, три переводных эквивалента многозначных слов переводимого предложения соответствуют одной микротеме: q_3 = *сеть* (переводной эквивалент многозначного слова web), q_6 = *интерфейс* (переводной эквивалент многозначного слова interface), q_{11}=*браузер* (пере-

водной эквивалент многозначного слова browser). При переводе предложения выбираются именно эти переводные эквиваленты, так как в одно предложение относится к одной микротеме, и все переводные эквиваленты должны относиться именно к этой микротеме. После проведения семантического анализа, позволившего снять многозначность переводимых слов предложения, получим следующий автоматический перевод предложения: *Некоторые общие клиентские интерфейсы сети — также известные, как сетевые браузеры — включающие NetScape, Internet Explorer и другие, проверяют электронные документы.*

Выводы

Таким образом, использование метода компараторной идентификации позволяет автоматически разделять элементы смысла многозначных слов переводимого предложения на тождественные (по отношению к определенной) микротемы. Применение данного метода на этапе семантического анализа в трансферных системах машинного перевода позволяет в ряде случаев снять семантическую омонимию многозначных слов предложения и тем самым уменьшить количество семантических ошибок переводимых текстов.

Список литературы: 1. *Мельчук И.А.* Опыт теории лингвистических моделей "Смысл-Текст".— М., 1974. —314 с. 2. *Апресян Ю.Д.* Избранные труды: Т. I: Лексическая семантика — М.: "Яз. рус. культуры", 1995. С. 472. 3. *Бондаренко М.Ф., Шабанов-Кушнаренко Ю.П.* Теория интеллекта: Учебник. — Харьков: ООО» Компания СМИТ», 2006.— 576 с. — На русск.языке. 4 .*Schneiderman R.A.* Why librarians should rule the net // E-NODE.— 1996.—Vol.1, N 4.— 5, Septem.

Поступила в редколлегию 01.10.2007